GigaScience

The Gene Expression Deconvolution Interactive Tool (GEDIT): Accurate Cell Type Quantification from Gene Expression Data

--Manuscript Draft--

Manuscript Number:	GIGA-D-20-00251R1
Full Title:	The Gene Expression Deconvolution Interactive Tool (GEDIT): Accurate Cell Type Quantification from Gene Expression Data
Article Type:	Research
Funding Information:	Biomedical Big Data Training Grant (5T32LM012424-03) — Dr. Brian B. Nadel
Abstract:	The cell type composition of heterogeneous tissue samples can be a critical variable in both clinical and laboratory settings. However, current experimental methods of cell type quantification (e.g. cell flow cytometry) are costly, time consuming, and can introduce bias. Computational approaches that infer cell type abundance from expression data offer an alternate solution. While these methods have gained popularity, most are limited to predicting hematopoietic cell types and do not produce accurate predictions for stromal cell types. Many of these methods are also limited to particular platforms, whether RNA-seq or specific microarrays. We present the Gene Expression Deconvolution Interactive Tool (GEDIT), a tool that overcomes these limitations, compares favorably with existing methods, and provides superior versatility. Using both simulated and experimental data, we extensively evaluate the performance of GEDIT and demonstrate that it returns robust results under a wide variety of conditions. These conditions include a variety of platforms (microarray and RNA-seq), tissue types (blood and stromal), and species (human and mouse). Finally, we provide reference data from eight sources spanning a wide variety of stromal and hematopoietic types in both human and mouse. This reference database allows the user to obtain estimates for a wide variety of tissue samples without having to provide their own data. GEDIT also accepts user submitted reference data, thus allowing the estimation of any cell type or subtype, provided that reference data is available.
Corresponding Author:	Brian Nadel University of California Los Angeles SANTA MONICA, CA UNITED STATES
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	University of California Los Angeles
Corresponding Author's Secondary Institution:	
First Author:	Brian B. Nadel
First Author Secondary Information:	
Order of Authors:	Brian B. Nadel David Lopez Dennis J. Montoya Feiyang Ma Hannah Waddel Misha M. Khan Serghei Mangul Matteo Pellegrini
Order of Authors Secondary Information:	
Response to Reviewers:	This following text is also included in the uploaded file "Response Letter":

Dear Giga Science Reviewer

Reviewer reports:

Reviewer #1: The authors present an algorithm called GEDIT using information from a reference dataset to estimate cell type abundances in a target dataset. This manuscript is not qualified to be published in the Giga Science because of the following reasons:
1. GEDIT does not show enough novelty.

As the authors stated in the response document, GEDIT has two key innovations: signature gene selection by information entropy and the row scaling step. As shown in Figure 2a, the signature gene selection only has a limited improvement than the others in terms of error. The authors also did not have enough evidence to support how and why the row scaling step is helpful. On top of these 2 data preprocessing steps, I did not find any innovations on the model of a non-negative linear regression.
We refer the reviewer to figure 4D, in which we demonstrate that row scaling dramatically improves accuracy of predictions. Here, we vary the row scaling parameter in order to evaluate the effect of row scaling on accuracy. When this parameter is set to 0 (the default value for GEDIT) both average and maximum error are greatly reduced compared to a setting of 1.0 (i.e. row scaling is disabled).

2. GEDIT does not add much values in the field.
Although GEDIT is shown to have appealing results in comparison to existing methods on benchmarking experiments, it doesn't significantly outperform other methods in real data analysis in regards to Pearson correlation and average error. As demonstrated in Supplementary Figures 4 and 5, a main determinant of results quality for deconvolution is the reference used, not necessarily the algorithm.

We refer to Figure 4, which demonstrates that GEDIT does indeed outperform other tools in terms of Pearson correlation. When any of the four references are used, GEDIT produces higher correlations between predicted and actual fractions than any other tool (the leftmost "all" column). In addition, the highest observed correlations for each of the 3 mixtures sets (ascites, cellmix, blood) are achieved by GEDIT (i.e. when the Human Primary Cell Atlas, LM22, and BLUEPRINT are used as reference sources, respectively). Lastly, unlike the three other reference-based tools, GEDIT produces positive correlations for all cell types regardless of choice of reference. GEDIT similarly outperforms other tools when evaluated by error (Supplementary Figure 2).
Moreover, we include in the supplementary materials an additional comparison of bulk deconvolution tools (Supplementary Figure 6). Here, we compare the error of CIBERSORT, DeconRNASeq, dtangle, and GEDIT when applied to simulated pancreatic islet mixtures. Again, GEDIT outperforms the other deconvolution tools.
In addition, we are pursuing a separate project performing extensive benchmarking of the current field of deconvolution tools. We believe more comprehensive comparisons of these tools' performances is appropriate, but that such an undertaking represents a separate project that should not be bundled with the publication of a new tool. For the interest of the editors and reviewers, we attach an early version of this benchmarking manuscript.

3. I still think comparing GEDIT to other methods using single-cell RNA-seq as a reference is necessary.
Utilizing single-cell RNA-seq for deconvolution becomes cutting-edge research and many tools have been designed for this purpose such as MuSic. These methods are proved to have superior performance using microarray as references and are commonly used. It is critical to demonstrate whether GEDIT has better performance than these methods.

At the reviewer's request, we have performed a comparison between GEDIT and two well known single cell deconvolution tools (SCDC and MuSiC). Specifically, we utilize the testing framework developed by the SCDC authors to prepare synthetic mixtures using single cell data from pancreatic islets. We include an additional section describing these results in the main manuscript (text below), and also refer the reviewer to Figure 5 and Supplementary Figure 5.

Comparison to Single Cell Methods
We also compare GEDIT to two contemporary deconvolution tools that utilize single cell data as their reference, namely SCDC and MuSiC [10,11]. We reproduce the steps provided by the SCDC authors to generate two sets of 100 simulated pancreatic mixtures. These data are created in silico using single cell data from two recent studies, and contain randomized mixtures of alpha, beta, delta, and gamma cells from pancreatic islets [33,34]. Data from a third study was used as a reference for all 3 tools, and similarly contains alpha, beta, gamma, and delta cells [35]. In the case of SCDC and MuSiC, these data are used in their original single cell form. For GEDIT, pseudo-bulk expression profiles for each of the four cell types were created by averaging the expression values of each member cell (e.g. expression of all alpha cells were averaged to create an alpha cell reference profile).
The results of GEDIT compare favorably to the two single cell tools (Figure 5). GEDIT produces the lowest error on the two sets of simulated mixtures by a significant margin. Based on the metric of correlation between predicted and actual fractions, GEDIT produces results comparable to SCDC, and either comparable or superior to MuSiC, depending on the set of mixtures (Table 4, Supplementary Figure 5). Thus, by using the methodology of averaging cell clusters in the reference dataset, GEDIT can be applied to datasets suitable for SCDC or MuSiC. We also apply three other bulk deconvolution tools to this same dataset, and show that GEDIT provides the best performance out of the four (Supplementary Figure 6).

Reviewer #2: For the most part the responses are sufficient and the authors have addressed the concerns, and the manuscript is improved. I especially appreciate that claims have been toned down and better contextualized.

A small issue remains about minor comment 6. My point was that the readers should be made clearly aware that doing three deconvolutions is not ideal, and strictly speaking invalid (e.g. cell contents totalling over 100%). I suggested two hypothetical ways to avoid this and am not at all surprised it's not easy to fix with data on hand. In the context of the demonstration in this particular study, what the authors did originally is acceptable. The problem I raised is if people start copying that practice in their own studies, and in the authors' own interest they presumably wouldn't want to be seen as endorsing it. The statement in the supplement that "creating a comprehensive reference from single cell data will likely produce superior results" should be more prominent, and it's not just about superior results, it's also about validity of having a single reference vs. multiple independent deconvolutions.

To be concrete, I'd suggest that in the main body a parenthetical could be added to the effect that "it would be more appropriate to have a single reference containing all cell types and performing a single deconvolution; see supplement for discussion". As it stands I don't think the addition to the supplement is referenced in the main paper.

We appreciate this feedback. We have adjusted the language in this section as follows:

To assess the use of GEDIT across very large datasets, we applied the tool to 17,382 GTEx RNA-seq samples collected from various tissues. However, no single reference contained all cell types expected to be present and combining references from separate experiments and platforms is problematic (Supplementary Figures 9-11). Therefore, we took an alternate approach by performing deconvolution three times using three separate references (BlueCode, Human Primary Cell Atlas, Skin Signatures). We then combine these outputs by taking their median value; after normalization, we treat this median value as a final cell type estimate (see Supplementary Materials for more details). While this approach did enable predictions spanning a larger number of cell types than are present in any one reference matrix, it must me noted that it is not a proper substitute for a single unified reference (Figure 8).

Additional Information:	
Question	**Response**
Are you submitting this manuscript to a special series or article collection?	No

Experimental design and statistics Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends. Have you included all the information requested in your manuscript?	Yes
Resources A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible. Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?	Yes
Availability of data and materials All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in publicly available repositories (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript. Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist?	Yes

The Gene Expression Deconvolution Interactive Tool (GEDIT): Accurate Cell Type Quantification from Gene Expression Data

Brian B. Nadel*[1], David Lopez[1], Dennis J. Montoya[1], Feiyang Ma[1], Hannah Waddel[3], Misha M. Khan[4], Serghei Mangul[5], Matteo Pellegrini[1,2]

[1]Bioinformatics Interdepartmental Degree Program, Molecular Biology Institute, Department of Molecular Cellular and Developmental Biology, and Institute for Genomics and Proteomics, University of California Los Angeles, Los Angeles, CA

[2]Department of Dermatology, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA

[3]Department of Mathematics, University of Utah, Salt Lake City, UT

[4]Departments of Biology and Computer Science, Swarthmore College, Swarthmore, PA

[5] Department of Clinical Pharmacy, USC School of Pharmacy, Los Angeles, CA

*Corresponding author. brian.nadel@gmail.com. Phone 310-963-7077.

ORCID IDs: **BBN** 0000-0002-9583-8169, **DJM** 0000-0002-0802-3309, **FM** 0000-0002-6260-0787, **HW** 0000-0001-5970-2750, **MMK** 0000-0002-6135-5580, **SM** 0000-0003-4770-3443, **MP** 0000-0001-9355-9564

Abstract

The cell type composition of heterogeneous tissue samples can be a critical variable in both clinical and laboratory settings. However, current experimental methods of cell type quantification (e.g. cell flow cytometry) are costly, time consuming, and can introduce bias. Computational approaches that infer cell type abundance from expression data offer an alternate solution. While these methods have gained popularity, most fail to produce accurate predictions for the full range of platforms currently used by researchers, or for the wide variety of tissue types often studied. We present the Gene Expression Deconvolution Interactive Tool (GEDIT), a tool that overcomes these limitations, compares favorably with existing methods, and provides superior versatility. Using both simulated and experimental data, we extensively evaluate the performance of GEDIT and demonstrate that it returns robust results under a wide variety of conditions. These conditions include multiple platforms (microarray and RNA-seq), tissue types (blood and stromal), and species (human and mouse). Finally, we provide reference data from eight sources spanning a broad range of stromal and hematopoietic types in both human and mouse. This reference database allows the user to obtain estimates for a wide variety of tissue samples without having to provide their own data. GEDIT also accepts user submitted reference data, thus allowing the estimation of any cell type or subtype, provided that reference data is available.

Author Summary

The Gene Expression Deconvolution Interactive Tool (GEDIT) is a robust and accurate tool that uses gene expression data to estimate cell type abundances. Extensive testing on a variety of tissue types and technological platforms demonstrates that GEDIT provides greater versatility than other cell type deconvolution tools. GEDIT utilizes reference data describing the expression profile of purified cell types, and we provide in the software package a library of reference matrices from various sources. GEDIT is also flexible and allows the user to supply custom reference matrices. A GUI interface for GEDIT are available online, as well as source code and reference matrices, are available

online [40,41].

Introduction

Cell type composition is an important variable in biological and medical research. In laboratory experiments, cell sample heterogeneity can act as a confounding variable. Observed changes in gene expression may result from changes in the abundance of underlying cell populations, rather than changes in expression of any particular cell type [1]. In clinical applications, the cell type composition of tissue biopsies can inform treatment. For example, in cancer, the number and type of infiltrating immune cells has been shown to correlate highly with prognosis ([2], [3], [4]). Moreover, patients with a large number of infiltrating T cells are more likely to respond positively to immunotherapy [5].

For many years, cell flow cytometry via FACS sorting has been the standard method of cell type quantification. More recently, single cell RNA-seq methods such as 10x Chromium, Drop-Seq, and Seq-Well have become available [6,7]. However, both approaches suffer from significant limitations. FACS sorting is cumbersome and expensive, and some sample types require hours of highly skilled labor to generate data. Similarly, single cell RNA-seq methods remain expensive for large sample studies. Additionally, cell types such as neurons, myocytes, and adipocytes are difficult for these technologies to capture due to cell size and morphology.

Both FACS sorting and single cell methods have the potential to introduce bias, as these technologies require that tissue samples be dissociated into single cell suspensions. Many stromal cell types are tightly connected to one another in extracellular matrices. The procedures necessary to create single cell suspensions can damage some cells, while others remain in larger clusters that are not captured or sequenced. Consequently, subtle differences in sample preparation can produce dramatically different results [8,9]. While FACS sorting and single cell methods can produce pure samples of each cell type, the observed cell counts may not accurately represent the cell type abundances in the original sample. Tools like SCDC and MuSiC utilize single cell reference data to perform bulk deconvolution, but require that multi-subject single cell data be available for all the cell types of interest, which is not always the case [10,11].

During the past several years, digital means of cell type quantification, often referred to as cell type deconvolution or decomposition, have become a popular complement to FACS sorting and single cell approaches. However, these methods are still developing, and often suffer from limitations. For example, tools MCP-Counter and xCell allow for deconvolution of a set of predefined cell types, but do not support the inclusion of additional cell types or subtypes in a user friendly manner [12,13]. CIBERSORT is slow to run on large datasets, particularly if signature genes are not specified, and provides reference data only for hematopoietic cell types [14].

To overcome some of the limitations of existing cell abundance estimation tools, we present the Gene Expression Deconvolution Interactive Tool (GEDIT). GEDIT utilizes gene expression data to accurately predict cell type composition of tissue samples. We have assembled a library of reference data from 11 distinct sources and use these data to generate thousands of synthetic mixtures. In order to produce optimal results, these synthetic mixtures are used to test and refine the approaches and parameters used by GEDIT. We compare the performance of GEDIT relative to other tools using three sets of mixtures containing known cell type proportions: 12 *in vitro* mixtures of immune cells sequenced on microarrays, six RNA-seq samples collected from ovarian cancer ascites, and eight RNA-seq samples collected from blood. We also use GEDIT to deconvolute two sets of human tissue samples: 21 skin samples from patients with skin diseases, and 17,382 samples of varied tissues from the GTEx database. Lastly, we apply GEDIT to the Mouse Body Atlas, a collection of samples collected from various mouse tissues and cell types. We find that GEDIT compares favorably to other cell type deconvolution tools and is effective across a broad range of datasets and conditions.

Results

Reference Data

Reference data profiling the expression of purified cell types is a requirement for reference-based deconvolution. Methods that do not directly require reference data, such as non-negative matrix factorization, still require knowledge of expression profiles or marker genes in order to infer the identity of the predicted components. For this study, we have assembled or downloaded a set of 11 reference matrices, each containing the expression profiles of eight to 29 cell types (Table 1). These data sources span multiple platforms, including bulk RNA-seq, microarray, and single-cell RNA-seq. Complete details on the sources and assembly of these matrices are described in the methods [14–24].

Matrix	**Species**	**Reference**	**Platform**	**# of Cell Types**	**Cell Types**
Human Skin Signatures	Human	(Swindell et al. 2013)	Multi-Microarray	21	Immune
Human Body Atlas	Human	(Su et al. 2004)	Affymetrix U133A/GNF1H	13	Immune
Human Primary Cell Atlas	Human	(Mabbott et al. 2013)	Affymetrix U133 Plus 2.0	26	Immune and Stromal
BLUEPRINT*	Human	(Martens and Stunnenberg 2013)	Bulk RNA-Seq	8	Immune
ENCODE*	Human	(ENCODE Project Consortium 2004)	Bulk RNA-Seq	29	Mostly Stromal
LM22	Human	(Newman et al. 2015)	Affymetrix Microarray	22	Immune
10x Single Cell Dataset*	Human	(Zheng et al. 2017)	Single Cell RNA-Seq	9	Immune
ImmunoStates	Human	(Vallania et. al., 2018)	Multi-Microarray	20	Immune

Tabula Muris	Mouse	(The Tabula Muris Consortium, 2018)	Single Cell RNA-seq	12	Immune and Stromal
Mouse Body Atlas	Mouse	(Lattin et al, 2008)	Affymetrix Mouse Genome 430 2.0 Array	20	Immune and Stromal
ImmGen	Mouse	(Heng et al, 2008)	Affymetrix Gene 1.0 ST	137	Immune with many subtypes

Table 1. Library of Reference Data. Asterisk denotes matrices assembled from source data as part of this project. All matrices are compatible with GEDIT and available on the GitHub repository [40].

GEDIT Algorithm

GEDIT requires as input two matrices of expression values. The first is expression data is collected from the mixtures that will be deconvoluted; each column represents one mixture, and each row corresponds to a gene. The second matrix contains reference data, with each column representing a purified reference profile and each row corresponding to a gene. In a multi-step process, GEDIT utilizes the reference profiles to predict the cell type proportions of each submitted mixture (Figure 1).

Figure 1. The GEDIT pipeline. The input matrices are quantile normalized then reduced to matrices containing only signature genes. Next, a row-scaling step serves to control for the dominating effect of highly expressed genes. Lastly, linear regression is performed, and predictions of cell type abundances are reported to the user.

Input	**Description**	**Allowed Values**	**Default Value**

RefMat	Matrix of purified cell types	N by M matrix; N is number of genes, M is number of cell types	NA
MixMat	Matrix of mixtures to be deconvoluted	N by P matrix; N is number of genes, P is number of mixtures	NA
SigMeth	Method of signature gene selection	Entropy, MeanRat, MeanDiff, ZScore, fsRat, fsDiff	Entropy
NumSigs	Average number of signature genes per cell type	[1, 10,000]	50
MinSigs	Minimum number of signatures per cell type	[1,NumSigs]	=NumSigs
RowScale	Extent of per-row normalization	[0.0,1.0]	0

Table 2. GEDIT inputs include two matrices and four parameter settings. RefMat is an expression matrix documenting the expression profiles of each cell type to be estimated. MixMat is an expression matrix documenting expression values for each sample to be deconvoluted. SigMeth determines the method by which signature genes are selected. NumSigs determines the total number of signature genes, whereas MinSigs sets the minimum number of signature genes for each cell type. RowScale refers to the extent to which expression vectors are transformed to lessen the dominating effect of highly expressed genes, with a value of 0.0 representing the most extreme transformation. Default values were determined by evaluating performance on a set of synthetic mixtures (Figure 2).

Parameter Tuning

We generated a large number of synthetic mixtures *in silico* to test the efficacy of GEDIT and to assess how accuracy varies as a function of four parameter choices (SigMeth, NumSigs, MinSigs, RowScale, described in Table 2). We produced a total of 10,000 simulated mixtures of known proportions using data from four reference matrices: BLUEPRINT, The Human Primary Cell Atlas, 10x

Single Cell, and Skin Signatures. We then ran GEDIT on these simulated mixtures and evaluated its performance while varying four parameter settings (Figure 2) and other design choices. For this reason, these synthetic mixtures were not used to evaluate the performance of GEDIT relative to other tools. Instead, separate datasets were used for that purpose, as described in the section “Performance Comparison to Other Deconvolution Tools”. Based on these results, we selected default values for each parameter (SigMeth = Entropy, NumSigs = 50, MinSigs = 50, RowScale = 0.0). Full details on the generation of these simulations are described in the supplementary materials.

Figure 2. Effect of GEDIT parameter choices on accuracy of predictions in simulated experiments. 10,000 simulated mixtures were generated, each using one of four reference matrices, with either four, five, six, or ten cell types being simulated. Deconvolution was performed using a separate expression matrix than the one used to generate the mixtures. When not otherwise noted, we use the following parameters: signature selection method = entropy; number of signatures = 50; row scaling = 0.0; and number of fixed genes = number of signatures.

Preprocessing and Quantile Normalization

The first step in the GEDIT pipeline is to render the two matrices comparable. This is done by first excluding all genes that are not shared between the two matrices. Genes that have no detected expression in any reference cell type are also excluded, as they contain no useful information for deconvolution. Both matrices are then quantile normalized, such that each column follows the same distribution as every other; this target distribution is the starting distribution of the entire reference matrix.

Signature Gene Selection

GEDIT next identifies signature genes. Gene expression experiments can simultaneously measure tens of thousands of genes, but many of these genes are uninformative for deconvolution. Specifically, genes with similar expression levels across all cell types are of little use, as observed expression values in the mixtures offer no insight into cell frequencies. Genes that are highly expressed in a subset of cell types are more informative, and we refer to these as signature genes. By using only signature genes, rather than the entire expression matrix, the problem of deconvolution becomes more tractable and less computationally intensive. Moreover, identification of signature genes can be valuable to researchers for other applications (e.g. cell type assignment for scRNA-seq data).

In order to identify the best signature genes in a given reference matrix, GEDIT calculates a signature score for each gene. By default, this score is computed using the concept of information entropy. Information entropy quantifies the amount of information in a probability distribution, with highly uniform distributions having the highest entropy. The expression vector for each gene (i.e. the set of expression values across all cell types in the reference) is divided by its sum, such that the entries can be interpreted as probabilities. Information entropy is then calculated according to its mathematical definition (see Methods), and genes with the lowest entropy are selected as signature genes. Entropy is minimized when expression is detected only in a single cell type and maximized when expression values are equal across all cell types. Thus, by selecting genes with low entropy, we favor genes that are expressed in a cell type specific manner. By default, 50 signature genes are selected for each cell type in the reference matrix. We chose 50 signature genes, and entropy as our scoring method, because it returned optimal results when run on 10,000 synthetic mixtures (see Figure 2a,b).

We also evaluated the effect of accepting more signature genes for some cell types than others, depending on how many genes have low entropy. In this scheme, on average 50 signature genes are used per cell type. However, a fourth parameter is used, which specifies the minimum number of signature genes per cell type. After these have been selected, remaining signature genes are added based only on lowest entropy, regardless of cell type of maximal expression. We found that this parameter had minimal effect on accuracy, when applied to synthetic mixtures (Figure 2c). Therefore, this option is not used by default, though it can be specified by the user.

Row Scaling and Linear Regression

One complication in the application of linear regression to gene expression data is the drastically different scale at which some genes are expressed. For example, CD14 and THEMIS (Figure 3) have both been identified as signature genes: CD14 for monocytes and THEMIS for CD4+ T cells. However, CD14 is expressed at much higher levels in most cell types and will have a larger impact on the estimation of cell type composition, relative to THEMIS. In other words, the possible penalty resulting from a poor fit of CD14 is much larger than the penalty from a poor fit of THEMIS.

Figure 3. The “row scaling” transformation, as implemented by GEDIT. CD14 and THEMIS are two examples of signature genes with drastically different magnitudes of expression. CD14 is a signature gene for monocytes, and THEMIS for CD4+ T cells. The original expression vectors are transformed, such that all values fall between 0.0 and 1.0, equalizing the effect of genes with varying magnitudes of expression.

In order to equalize the effect of each signature gene on the linear regression, we implement a transformation that we term row scaling. Specifically, the range of all observed values for a particular gene (including reference cell types and samples) is adjusted such that the maximum value is 1.0 and the minimum value is 0.0. As a result, all genes have a comparable influence on the calculation of the linear regression solution, regardless of overall magnitude of expression. This transformation can be modulated by adjusting the row scaling parameter. By default, the value of this parameter is 0.0, and the transformation is applied as described above. Values between 0.0 and 1.0 are also allowed, which reduces the extent of the transformation (see Methods for details). Linear regression is then performed in R using the glmnet package, as described in the methods.

Performance Comparison to Other Deconvolution Tools

In order to assess the performance of GEDIT relative to other tools, we perform an experiment comparing GEDIT to 4 other deconvolution tools on datasets of known cell-type content

(CIBERSORT, DeconRNASeq, dtangle and xCell; [13,14,25,26]). Non-deconvolution tools like MCP-counter, SAVANT, and the DCQ algorithm are excluded from this study because they do not predict cell type fractions [12,27,28].Tools that require single cell data, such as MuSiC and CPM, are also excluded, as this study is limited to tools that operate on bulk expression data [11,29]. See Table 3 for a summary of current bulk deconvolution methods.

					Reference Data Provided with Tool		
Tool	**Publication**	**Custom Reference**	**Approach**	**Output**	**Number of Datasets**	**Cell Types**	**Species**
GEDIT	Nadel et. al., 2020	Yes	Deconvolution	Predicted Fractions	11	Immune and Stromal	Human, Mouse
CIBERSORT	Newman et. al., 2015	Yes	Deconvolution	Predicted Fractions	1	Immune	Human
xCell	Aran et. al., 2017	No	Marker Genes	Predicted Fractions	5	Immune and Stromal	Human
dtangle	Hunt et. al., 2018	Yes, if marker genes specified	Deconvolution	Predicted Fractions	0	N/A	N/A
DeconRNASeq	Gong et. al., 2013	Yes	Deconvolution	Predicted Fractions	0	N/A	N/A
DCQ/ImmQuant	Altboum et. al., 2014; Frishberg et. al., 2016	Yes	Deconvolution	Scores	3	Immune	Human, Mouse
CIBERSORT (absolute mode)	Newman et. al., 2015	Yes	Deconvolution	Scores	1	Immune	Human
SaVant	Lopez et. al., 2017	Yes, if marker genes specified	Marker Genes	Scores	12	Immune and Stromal	Human, Mouse
MCP-Counter	Becht et. al., 2016	No	Marker Genes	Scores	N/A	Immune and Stromal	Human

Table 3. High level characteristics of current cell type estimation tools. Some tools accept custom references, which allows the tool to estimate the abundance of cell types not present in the default reference. Tools listed here take one of two approaches: they either perform deconvolution (most commonly regression) or calculate a score based on intensity of marker gene expression. Depending

on the tool, the output can be interpreted as fractions corresponding to the abundance of each cell type, or as scores for each cell type that cannot necessarily be compared in an inter-cellular manner.

To perform this study, we utilize three datasets for which cell type fractions have been estimated using orthogonal methods. Two of these datasets were used in a recent benchmarking study [30]. Both are profiled using RNA-seq, and represent samples collected either from human cancer ascites or human blood [31,32]. In both cases, cell type fractions have been evaluated by FACS sorting. The final dataset was prepared *in vitro* and consists of six cell types that were physically mixed together (in known proportions) to prepare 12 mixtures. These mixtures were then profiled using an Illumina HT12 BeadChip microarray. Adding to the previous benchmarking study, we also explore the effect of using four separate reference datasets: The Human Primary Cell Atlas, LM22, ImmunoStates, and a reference constructed from BLUEPRINT data. For each dataset, all tools (except xCell) were run four times, each time using a different reference matrix.

The optimal choice of reference matrix varies greatly depending on the exact combination of tool, dataset, and cell type. While using LM22 often produces the most accurate results, there are many exceptions. For instance, DeconRNASeq and GEDIT produce their best results for the blood dataset when using the BLUEPRINT reference. For the ascites data, several tools prefer ImmunoStates as the optimal reference choice. The best choice of reference is highly dependent on the nature of the input data and on the tool being used. In practice, researchers may wish to perform deconvolution multiple times--in each case using a separate reference matrix--and compare results for consistency.

Compared to the other tools, GEDIT produces robust and consistently accurate results (Figure 4, Supplementary Figures 1-2). For many tools, the quality of predictions varies greatly depending on the cell type, dataset, or choice of reference matrix. When results are averaged across the four possible reference choices, GEDIT produces the minimum error and maximum correlation for all three datasets. This result suggests that GEDIT is a strong choice when researchers are using novel references matrices that have not been curated or tested.

Figure 4. Performance of five deconvolution tools when applied to a set of 26 physical samples from three sources. Actual cell type fractions are either known due to controlled cell mixing (Cell Mix) or estimated by FACS sorting (Ascites and Blood). In each instance, we calculate the correlation between actual cell type fractions and those predicted by deconvolution; deeper blues represent higher correlations (A). We similarly calculate average error, with deeper reds representing higher error (B). We test four different reference datasets for each tool, and averaged correlations across these 5 cases are shown in boxes. We calculate correlations for each cell type (right 5 columns), for each of the 3 mixtures (middle 3 columns), and for all predictions regardless of cell type or data source.

Lastly, we perform an evaluation of runtime required for each tool. We randomly select batches of 100, 200, 500, 1000, and 2000 samples from the GTEx database, and measure CPU time required to deconvolute these batches for each tool. The runtimes of GEDIT, dtangle, and DeconRNASeq scale well with growing input size, taking at most 20 minutes (Supplementary Figure 3). For larger input sizes, CIBERSORT can take over an hour.

Comparison to Single Cell Methods

We also compare GEDIT to two contemporary deconvolution tools that utilize single cell data as their reference, namely SCDC and MuSiC [10,11]. We reproduce the steps provided by the SCDC authors to generate two sets of 100 simulated pancreatic mixtures. These data are created *in silico* using single cell data from two recent studies, and contain randomized mixtures of alpha, beta, delta, and gamma cells from pancreatic islets [33,34]. Data from a third study was used as a reference for all 3 tools, and similarly contains alpha, beta, gamma, and delta cells [35]. In the case of SCDC and MuSiC, these data are used in their original single cell form. For GEDIT, pseudo-bulk expression profiles for each of the four cell types were created by averaging the expression values of each member cell (e.g. expression of all alpha cells were averaged to create an alpha cell reference profile).

The results of GEDIT compare favorably to the two single cell tools (Figure 5). GEDIT produces the lowest error on the two sets of simulated mixtures by a substantial margin. Based on the metric of correlation between predicted and actual fractions, GEDIT produces results comparable to SCDC, and either comparable or superior to MuSiC, depending on the set of mixtures (Figure 5c, Supplementary Figure 4). Thus, by using the methodology of averaging cell clusters in the reference dataset, GEDIT can be applied to datasets suitable for SCDC or MuSiC. We also apply three other bulk deconvolution tools to this same dataset and show that GEDIT provides the best performance out of the four (Supplementary Figure 5).

Figure 5. Performance of GEDIT and two single cell deconvolution tools when applied to simulated pancreatic islet mixtures. Two sets of 100 synthetic mixtures were used, each set from a separate data source of scRNA-seq alpha, beta, gamma, and delta cells [33,34]. Protocols developed by the SCDC authors were used to create synthetic mixtures *in silico*. Data from a third source was used as a reference [35].(A,B) Error distributions plots for each of the two datasets. Each point on a graph represents the percentage of predictions (y-axis) that are accurate within a particular error range (x-axis). (C) Summary statistics of predictions for each tool and dataset. Pearson correlation and average error between predicted and actual cell type fractions are shown

Skin Expression Data

We further validate GEDIT by using it to deconvolute a set of skin biopsies from humans with a variety of skin diseases [13]. The exact cell type composition of these samples is unknown, but we have reasonable expectations based on skin and disease biology. For example, macrophages are known to be abundant in granulomas of leprosy lesions, and Steven-Johnson Syndrome produces blisters that fill with large numbers of monocytes [36,37]. We find that, in all cases, predictions made by GEDIT conform well with these biological expectations. Keratinocytes are highly predicted in most cases, as one would expect with skin samples (Figure 6). Deviations from this pattern correspond with disease biology. Monocytes are highly predicted in Stevens-Johnson syndrome, as are macrophages in the three leprosy samples, and T cells in the Mycosis Fungoides (T cell lymphoma) sample. Three other deconvolution tools were also applied to this dataset, and predictions follow similar patterns (Supplementary Figure 6).

Figure 6. GEDIT predictions for 21 samples of various skin diseases. GEDIT correctly identifies keratinocytes and subcutaneous adipose as the most common cell. Deviations from this pattern correspond to disease biology. SJS represents blister fluid from Steven Johnson Syndrome, and is predominantly immune cells. LL and RR represent two forms of leprosy, which result in large numbers of macrophages. MF is a T Cell Lymphoma.

Application of GEDIT to Mouse Data

GEDIT can be used to decompose data from any organism for which reference data is available. Here, we demonstrate the efficacy of GEDIT when applied to the Mouse Body Atlas, a collection of tissue and cell type samples collected from mice (GEO:GSE10246) [23]. As reference data, we assembled a matrix of 12 cell types using single cell data from the Tabula Muris [22]. GEDIT correctly infers the identity of purified cell types, including six samples that consist of either pure NK cells, B cells, T cells, or granulocytes (Figure 7). An entry for macrophages is not available in the reference used, but most macrophage samples are identified as monocytes, which is the most similar cell type present in the reference matrix. For more complex tissues, GEDIT predicts cell type fractions that correspond to the biology of the samples. Hepatocytes are predicted to be highly prevalent in the liver sample (84%) and are not predicted in any other sample (less than 5% in all cases). Similar patterns hold for keratinocytes in the epidermis, epithelial cells in two intestinal samples and cardiac muscle cells in heart and muscle samples. Three other deconvolution tools were also applied to this dataset, and predictions follow similar patterns (Supplementary Figure 7).

Figure 7. GEDIT predictions on 30 samples collected from various mouse tissues and cell types (Mouse Body Atlas [23]). Predictions largely conform with tissue and cell biology.

Deconvolution of GTEx Database

To assess the use of GEDIT across very large datasets, we applied the tool to 17,382 GTEx RNA-seq samples collected from various tissues and accessed via the GTEx portal [38]. However, no single reference contained all cell types expected to be present and combining references from separate experiments and platforms is problematic (Supplementary Figures 8-10). Therefore, we took an alternate approach by performing deconvolution three times using three separate references (BlueCode, Human Primary Cell Atlas, Skin Signatures). We then combine these outputs by taking their median value; after normalization, we treat this median value as a final cell type estimate (see Supplementary Materials for more details). While this approach did enable predictions spanning a larger number of cell types than are present in any one reference matrix, it must be noted that it is not a proper substitute for a single unified reference (Figure 8).

Figure 8. GEDIT cell type predictions when applied to 17,382 samples from the GTEx database. Here, predictions have been averaged for each tissue of origin (see supplementary methods for additional details).

These predictions largely conform to biological expectations. For example, immune cells are predicted to have high abundance in blood and spleen, adipocytes in adipose tissue, Shwann cells in nerve and heart, and keratinocytes in skin. Each of these patterns matches expectations of which cell types should be present in these tissues. However, neither cardiac myocytes nor smooth muscle are highly abundant in GTEx muscle samples. This is likely because the GTEx samples are collected from skeletal muscle, which is known to have an expression profile that is distinct from that of cardiac and smooth muscle.

GEDIT Availability

GEDIT can be run online and source code, associated data, and relevant files are available via GitHub [40]. We provide access to the tool, a set of varied reference data, and two sample mixture matrices. The website automatically produces a heatmap of predicted proportions for the user, as well as a .tsv file. The user also has access to the parameter choices of GEDIT (signature gene selection method, number of signature genes, row scaling).

Methods

GEDIT Algorithm

Signature Gene Selection

During signature gene selection, we automatically exclude genes with zero detected expression in half or more of cell types. Observed expression values of exactly zero are often the result of either technical artefacts or resolution issues. Using such genes as signatures can result in inaccurate and highly unstable results, particularly when working with scRNA-seq derived data. As an additional safeguard, we treat all remaining expression values of zero as the lowest observed non-zero value in the matrix. Implementing this change has minimal effect on most genes but prevents genes with resolution issues from achieving artificially high scores. We consider this transformation valid, since values of zero generally do not represent zero expression, but rather an expression level below the detection limit of the technology used.

For any given gene, a scoring method takes as input the vector of the expression values across all reference cell types, and outputs a score. A gene is considered a potential signature gene in cell type X if it is expressed more highly in X than any other cell type. For each cell type, we keep only the N genes with the highest scores, where N is the NumSigs parameter.

Information entropy (H) is calculated using the following formula:

$$H = -\sum_{1}^{i} \quad [p_i * (p_i)\,] \qquad (1)$$

where p_i is the probability of the i^{th} observation. To apply this to expression values, we convert the vector of expression values into a vector of probabilities by dividing by its sum. In an equal mixture of each cell type, the i^{th} probability can be interpreted as the fraction of transcripts originating from the i^{th} cell type.

Row Scaling

During this step, we apply a transformation on the expression values for each gene. Each gene has measured expression in N purified cell types and M samples. Each of these values, X_{old}, is transformed according to the following formula:

$$X_{new} = (X_{old} - Min)/(Max - Min) * Max^{p} \quad (2)$$

Where Min is the minimum of all M + N original values, Max is the maximum of those values, and p is a tunable parameter with natural range p $\in$ [0.0,1.0]. This procedure produces values between the range of 0.0 and Max^{p}.

Linear Regression:

Non-negative linear regression was performed using the glmnet package in R. The glmnet function is used with lower.limits=0, alpha=0, lambda=0, intercept=FALSE. These settings perform a linear regression where all coefficients are non-negative, and with no regularization and no intercept term.

Reference Data

BLUEPRINT Reference Dataset

35 gene counts files were downloaded from the BLUEPRINT database, all collected from venous blood [18]. This included entries for CD14-positive, CD16-negative classical monocytes (5 samples), CD38-negative naive B cells (1), CD4-positive, alpha-beta T cell (8), central memory CD4-positive, alpha-beta T cell (2), cytotoxic CD56-dim natural killer cell (2), macrophage (4), mature neutrophil (10), and memory B Cell (1). When two or more transcripts appeared for a single gene, the transcript with the highest average expression was selected and others were excluded. Genes with no detected expression in any sample were also excluded, and then each sample was quantile normalized. Samples generally clustered by cell type, but we excluded one CD4-positive alpha-beta T cell. Replicates for each cell type were then collapsed into a single entry by taking the median value for each gene.

ENCODE Reference Dataset

106 transcript quantification files were downloaded from the ENCODE database [19]. These included all RNA-seq experiments collected from adult primary cells, excluding four with warnings. Warnings indicated that three samples suffered from low replicate concordance and one sample from low read depth, and these samples were excluded. All samples were processed by the Gingeras Lab at Cold Spring Harbor and mapped to GRCH38.

The samples were quantile normalized and clustered. In cases where multiple transcripts were measured for a single gene, the expression of that gene was calculated as the sum of all transcripts. At this time, 18 additional samples were excluded as they did not cluster with their replicates. Based on sample descriptions and data clustering, we found that the remaining 88 samples represented 28 unique cell types. We produced an expression profile for each cell type by merging all samples of that cell type via median average. For example, a cluster of 19 samples were labelled as endothelial cells (collected from various body locations) and were merged into a single entry termed canonical endothelial cells. This dataset spans a wide range of stromal cell types (e.g. smooth muscle, fibroblast, epithelial), but contains only a single entry for blood cells, which are labelled mononuclear cells.

We also combined the ENCODE and BLUEPRINT reference matrices into a single reference matrix, which we call BlueCode. We combined, then quantile normalized, the columns of both matrices. Possible batch effects in this combined matrix have not been fully evaluated.

10x Reference Dataset

We obtained single cell expression data for nine varieties of immune cells from the 10x website [20]. This included at least 2446 cells for each cell type, and at least 7566 cells for all cells other than CD14 monocytes. For each cell type, expression values for all cells were mean averaged to form an expression profile.

Tabula Muris Reference Dataset

We downloaded from the Tabula Muris single cell data for 12 clusters of mouse cell types. For each cluster, we averaged all cells of that cluster to produce a reference profile for the corresponding cell type.

Other Reference Datasets

Other datasets used in this project were obtained from their corresponding publications or GEO repositories. This includes a reference matrix of human skin signatures, the Human Body Atlas, the Human Primary Cell Atlas, LM22, ImmunoStates, the Mouse Body Atlas, and ImmGen [14–17,21,23,24].

Skin Diseases Data

We obtained expression data from 21 skin biopsies, collected from human patients with a variety of skin diseases. These data originally came from a wide range of sources and platforms, and were compiled into a single dataset by previous work [39]**.**

GTEx Data

GTEx data for 17,382 samples were obtained from the GTEx portal [38]. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. We ran GEDIT on all samples three times, each time using a different reference matrix (BlueCode, the Human Primary Cell Atlas, and Skin Signatures). For each cell type, we calculated our initial estimate as the median estimate across the three sets of predictions (or fewer, if that cell type is missing from one to two of the reference matrices). Lastly, for each sample we divided the vector of predictions by its sum, such that the final predictions sum to 100%.

Multi-Tool Performance Evaluation

In Vitro Immune Cell Mixture

Combinations of six immune cells (Neutrophils, Monocytes, Natural Killer Cells, B cells, and CD4 and CD8 T Cells) were mixed together and sequenced using an Affymetrix array. Whole blood from healthy human donors was supplied with informed consent through a sample sharing agreement with the UCLA/CFAR Virology Core Lab (grant number 5P30 AI028697). CD4+ T cells, CD8+ T cells, B cells, and NK cells were isolated using Stem Cell Technologies (Vancouver, BC, Canada) RosetteSep negative selection. Neutrophils were positively selected through the EasySep approach, according to the manufacturer’s specifications. Cells were then counted by hemocytometer and added at defined percentages to a total cell count of two million cells to create six different mixtures. Subsequently cells were processed for RNA isolation by AllPrep DNA/RNA. Illumina HT12 BeadChip microarray was performed by the UCLA Neuroscience Genomics Core. Data was normalized by quantile normalization through R ‘normalize.quantiles’ function.

RNA-seq Mixtures Used for Tool Evaluation

We also obtained two datasets used in a recent benchmarking study [30]. The first dataset is composed of three RNA-seq samples, each with two technical replicates that represent biopsies of ovarian cancer ascites [32]. The second dataset is composed of RNA-seq collected from the blood of healthy individuals, some of whom recently received an influenza vaccine [31]. These data were downloaded from the GitHub for the benchmarking paper, which also contained FACS estimates for six cell types for the ascites data (B cells, dendritic cells, NK cells, T cells, macrophages, neutrophils) and five cell types for the blood data (B cells, dendritic cells, T cells, monocytes, natural killer cells). However, since dendritic cells were never present at more than 3.5% abundance, we did not evaluate performance for this cell type.

Tools

We installed and ran GEDIT, CIBERSORT, DeconRNASeq and dtangle on the hoffman2 computational cluster at UCLA. xCell was run using the online interface [13]. The default choice for genes signatures (xCell =64) was used. The RNA-seq option was selected for the 2 RNA-seq datasets (blood and ascites), but not for the *in vitro* dataset, which was sequenced on microarray.

xCell produces 67 output scores, seven of which were used in this study. These were the entries labelled “B-Cells”, “Macrophages”, “Monocytes”, “NK cells”, “Neutrophils”, “CD4+ T cells” and “CD8+ T Cells”. As suggested by the xCell authors, the outputs for CD4 and CD8 T cell subtypes were summed to produce a final output for total T cells.

Reference Data

We evaluated the performance of the four reference-based tools (GEDIT, CIBERSORT, DeconRNASeq and dtangle) using each of four choices of reference matrix (LM22, ImmunoStates, BLUEPRINT, and the Human Primary Cell Atlas). The BLUEPRINT and Human Primary Cell Atlas reference matrices differ from ImmunoStates and LM22 in that they contain tens of thousands of genes, many of which should not be considered signature genes. This contrasts to ImmunoStates and LM22; each reference matrix contains fewer than 600 genes, which have been specifically identified as signature genes by previous work [14,21]. We include both forms of reference matrices in order to evaluate the input requirements of the tools studied.

Depending on the choice of reference matrix, reference-based tools often produce multiple outputs for some cell types, each representing a cell subtype. This includes B cells (naïve and memory), Monocytes (CD14 and CD16), NK cells (resting and active) and T cells (many subtypes including varieties of CD4 and CD8). In each case, the outputs for each subtype were summed in order to produce a total score for each greater cell type.

Discussion

GEDIT is an expression-based cell type quantification tool that offers unique flexibility and accuracy in a wide variety of contexts. Using both simulated and experimental data, we demonstrate that GEDIT produces high-quality predictions for multiple platforms, species, and a diverse range of cell types, outperforming other tools in many cases. We include in the software package a comprehensive library of reference data, which facilitates application of GEDIT to a wide range of tissue types in both human and mouse. GEDIT can also accept reference data supplied by the user, which can be derived from bulk RNA-seq, scRNA-seq, or microarray experiments. GEDIT represents a competitive addition to the suite of existing tissue decomposition tools while maintaining flexibility and performance robustness.

As part of this project, we perform a study in which we compare the performance of several deconvolution tools using multiple metrics. Unlike previous evaluation studies, we explore the effect of reference choice by running tools multiple times with reference data from different sources. Choice of optimal reference has a large impact on the accuracy of many tools, but GEDIT provides robust performance and accurate estimates for many possible reference choices. While all efforts were taken to perform this comparison in an unbiased manner, the authors note that development of the tool was still underway when the first comparisons were made. All code and inputs used to reproduce this study are included in the GitHub repository [40], with the exception of CIBERSORT code, which is limited by copyright.

The high performance of GEDIT is due to two key innovations. Firstly, signature gene selection by information entropy serves to select genes that are the most informative for deconvolution. Secondly, the row scaling step, which aims to equally weight all signature genes into the final estimate, even those with comparatively low expression. In addition, the flexibility of GEDIT and the diverse set of reference matrices we provide allows GEDIT to be easily applied in a wide range of circumstances.

The output of GEDIT represents the fraction of mRNA originating from each cell type. This is an effective measure of the transcriptional contribution of each cell type in a mixture. However, in cases where some cell types consistently produce more or less mRNA per cell, this measure may not represent cell counts. Data capturing the average mRNA content per cell is becoming more widely available in the form of single cell experiments and could in principle be used to convert our fractions into cell counts.

When extensively applied to several large public datasets, GEDIT produces predicted cell type fractions that conform with biological expectations. When used to decompose skin biopsies, keratinocytes are found to be the most abundant cell type. Variations in the abundance of other cell types conform to expected immune responses across diseases. Similarly, cell type predictions of GTEx samples are concordant with our expectations of the dominant cell types across tissues. Schwann cells, keratinocytes, adipose cells, and immune cells are found to be most abundant in nerve, skin, adipose tissue, and blood, respectively.

Single cell RNA-seq is an emerging approach to study the composition of cell types within a sample. Due to biases associated with the capture of different cell types, these methods are not always capable of accurately quantifying cell type populations [8]. However, the pure reference profiles produced by existing methods can be used by GEDIT to generate accurate estimates of cell type populations. Thus, GEDIT circumvents some of the biases associated with the preparation of samples for both scRNA-seq and FACS. GEDIT is freely available, and therefore an extremely economical option for researchers, particularly those who profile expression data for other purposes.

GEDIT produces accurate results when tested on mixtures of human immune cells. Compared to other tools, GEDIT produces the lowest error in the majority of scenarios in the studied mixtures. GEDIT provides increased flexibility over previously developed tools, as we provide a set of reference matrices for varied cell types for both mouse and human datasets.

GEDIT provides unique advantages to researchers, especially in terms of cell type, species and platform flexibility, and constitutes a useful addition to the existing set of tools for tissue decomposition. Our efficient decomposition methodology has been extensively optimized and we find that it performs robustly across a broad range of tissues in both mouse and human datasets. Our future work will extend reference matrices to facilitate application of GEDIT on varied bulk gene expression datasets.

Availability of Source Code and Requirements

- Project name: GEDIT
- Project Home Page: https://github.com/BNadel/GEDIT
- Programming Languages: Python 2.0, R
- Other requirements: numpy, glmnet
- Operating Systems: Linux
- License: MIT

Availability of Data and Materials

All data used in this paper are freely available on GitHub [40], as well as their original sources. Code for DeconRNASeq was obtained as an R package from the CRAN repository. Code for CIBERSORT was obtained by requesting it via the web portal [14], and code for dtangle from the project's GitHub page [26].

Reference data is also available from their original sources. Most datasets can be found on project website pages or from public databases. These include BLUEPRINT [18], ENCODE [19], the

Human Primary Cell Atlas [17], LM22 [14], 10x Genomics [20], Tabula Muris [22], the Mouse Body Atlas [23], and ImmGen [24]. Some reference matrices were obtained as supplementary files from the publications listed in Table 1.

Expression values for the blood and ascites RNA-seq datasets were obtained from the GitHub repository and are also available at GEO: GSE64655. The *in vitro* mixture of immune cells was prepared by our lab, and available on our GitHub page. All supporting data and materials are available in the *GigaScience* GigaDB database [41].

Additional Files

Supplementary Materials and Figures. Further details regarding synthetic mixture generation, deconvolution tool comparisons, and applications to the skin, mouse, and GTEx datasets.

Abbreviations

GEDIT: Gene Expression Deconvolution Interactive Tool; FACS: fluorescence-activated cell sorting; CPM: Cell Population Mapping; DCQ: Digital Cell Quantifier; NK cell: natural killer cell; scRNA-seq: single cell RNA sequencing; RNA: ribonucleic acid; MCP-Counter: microenvironment cell population counter; SaVanT: Signature Visualization Tool.

Acknowledgments

We acknowledge the Biomedical Big Data Grant (5T32LM012424-03) for supporting Brian Nadel during the course of this research. We also acknowledge the Bruins-in-Genomics Summer Undergraduate Research Program for supporting Hannah Waddel and Misha Khan during the summer of 2017, when they contributed to this work. We also thank Lana Martin for her help with editing and proofreading the manuscript.

Competing Interests

The authors declare that they have no competing interests.

References

1. Bolen CR, Uduman M, Kleinstein SH. Cell subset prediction for blood genomic studies. BMC Bioinformatics. 2011;12: 258.

2. Gentles AJ, Newman AM, Liu CL, Bratman SV, Feng W, Kim D, et al. The prognostic landscape of genes and infiltrating immune cells across human cancers. Nat Med. 2015;21: 938–945.

3. Li B, Severson E, Pignon J-C, Zhao H, Li T, Novak J, et al. Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. Genome Biol. 2016;17: 174.

4. Fridman WH, Pagès F, Sautès-Fridman C, Galon J. The immune contexture in human tumours: impact on clinical outcome. Nat Rev Cancer. 2012;12: 298–306.

5. Şenbabaoğlu Y, Gejman RS, Winer AG, Liu M, Van Allen EM, de Velasco G, et al. Tumor immune microenvironment characterization in clear cell renal cell carcinoma identifies prognostic and immunotherapeutically relevant messenger RNA signatures. Genome Biol. 2016;17: 231.

6. Gierahn TM, Wadsworth MH, Hughes TK, Bryson BD, Butler A, Satija R, et al. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. Nature Methods. 2017. pp. 395–398. doi:10.1038/nmeth.4179

7. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell. 2015;161: 1202–1214.

8. Hines WC, Su Y, Kuhn I, Polyak K, Bissell MJ. Sorting out the FACS: a devil in the details. Cell Rep. 2014;6: 779–781.

9. Denisenko E, Guo BB, Jones M, Hou R, de Kock L, Lassmann T, et al. Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. Genome Biol. 2020;21: 130.

10. Dong M, Thennavan A, Urrutia E, Li Y, Perou CM, Zou F, et al. SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references. Brief Bioinform. 2020. doi:10.1093/bib/bbz166

11. Wang X, Park J, Susztak K, Zhang NR, Li M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nat Commun. 2019;10: 380.

12. Becht E, Giraldo NA, Lacroix L, Buttard B, Elarouci N, Petitprez F, et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. Genome Biol. 2016;17: 218.

13. Aran D, Hu Z, Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biol. 2017;18: 220.

14. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015;12: 453–457.

15. Swindell WR, Johnston A, Voorhees JJ, Elder JT, Gudjonsson JE. Dissecting the psoriasis transcriptome: inflammatory- and cytokine-driven gene expression in lesions from 163 patients. BMC Genomics. 2013;14: 527.

16. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, et al. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A. 2004;101: 6062–6067.

17. Mabbott NA, Baillie JK, Brown H, Freeman TC, Hume DA. An expression atlas of human primary cells: inference of gene function from coexpression networks. BMC Genomics. 2013;14: 632.

18. Martens JHA, Stunnenberg HG. BLUEPRINT: mapping human blood cell epigenomes.

Haematologica. 2013;98: 1487–1489.

19. Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, Hilton JA, Jain K, Baymuradov UK, Narayanan AK, Onate KC, Graham K, Miyasato SR, Dreszer TR, Strattan JS, Jolanki O, Tanaka FY, Cherry JM. The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. 2018 Jan 4;46(D1):D794-D801.

20. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8: 14049.

21. Vallania F, Tam A, Lofgren S, Schaffert S, Azad TD, Bongen E, et al. Leveraging heterogeneity across multiple datasets increases cell-mixture deconvolution accuracy and reduces biological and technical biases. Nat Commun. 2018;9: 4735.

22. Consortium TTM, The Tabula Muris Consortium, Coordination O, Coordination L, Organ collection and processing, Library preparation and sequencing, et al. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. Nature. 2018. pp. 367–372. doi:10.1038/s41586-018-0590-4.

23. Lattin JE, Schroder K, Su AI, Walker JR, Zhang J, Wiltshire T, et al. Expression analysis of G Protein-Coupled Receptors in mouse macrophages. Immunome Res. 2008;4: 5.

24. Heng TSP, Painter MW, Immunological Genome Project Consortium. The Immunological Genome Project: networks of gene expression in immune cells. Nat Immunol. 2008;9: 1091–1094.

25. Gong T, Szustakowski JD. DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. Bioinformatics. 2013;29: 1083–1085.

26. Hunt GJ, Freytag S, Bahlo M, Gagnon-Bartsch JA. dtangle: accurate and robust cell type deconvolution. Bioinformatics. 2018. doi:10.1093/bioinformatics/bty926.

27. Lopez D, Montoya D, Ambrose M, Lam L, Briscoe L, Adams C, et al. SaVanT: a web-based tool for the sample-level visualization of molecular signatures in gene expression profiles. BMC Genomics. 2017;18: 824.

28. Altboum Z, Steuerman Y, David E, Barnett-Itzhaki Z, Valadarsky L, Keren-Shaul H, et al. Digital cell quantification identifies global immune cell dynamics during influenza infection. Mol Syst Biol. 2014;10: 720.

29. Frishberg A, Peshes-Yaloz N, Cohn O, Rosentul D, Steuerman Y, Valadarsky L, et al. Cell composition analysis of bulk genomics using single-cell data. Nat Methods. 2019;16: 327–332.

30. Sturm G, Finotello F, Petitprez F, Zhang JD, Baumbach J, Fridman WH, et al. Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. Bioinformatics. 2019;35: i436–i445.

31. Hoek KL, Samir P, Howard LM, Niu X, Prasad N, Galassie A, et al. A Cell-Based Systems Biology Assessment of Human Blood to Monitor Immune Responses after Influenza Vaccination. PLOS ONE. 2015. p. e0118528. doi:10.1371/journal.pone.0118528

32. Schelker M, Feau S, Du J, Ranu N, Klipp E, MacBeath G, et al. Estimation of immune cell content in tumour tissue using single-cell RNA-seq data. Nat Commun. 2017;8: 2032.

33. Baron M, Veres A, Wolock SL, Faust AL, Gaujoux R, Vetere A, et al. A Single-Cell Transcriptomic

Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. Cell Syst. 2016;3: 346–360.e4.

34. Segerstolpe Å, Palasantza A, Eliasson P, Andersson E-M, Andréasson A-C, Sun X, et al. Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. Cell Metab. 2016;24: 593–607.

35. Xin Y, Kim J, Okamoto H, Ni M, Wei Y, Adler C, et al. RNA Sequencing of Single Human Islet Cells Reveals Type 2 Diabetes Genes. Cell Metab. 2016;24: 608–615.

36. de Sousa JR, Lucena Neto FD, Sotto MN, Quaresma JAS. Immunohistochemical characterization of the M4 macrophage population in leprosy skin lesions. BMC Infect Dis. 2018;18: 576.

37. Lin C-C, Chen C-B, Wang C-W, Hung S-I, Chung W-H. Stevens-Johnson syndrome and toxic epidermal necrolysis: risk factors, causality assessment and potential prevention strategies. Expert Rev Clin Immunol. 2020;16: 373–387.

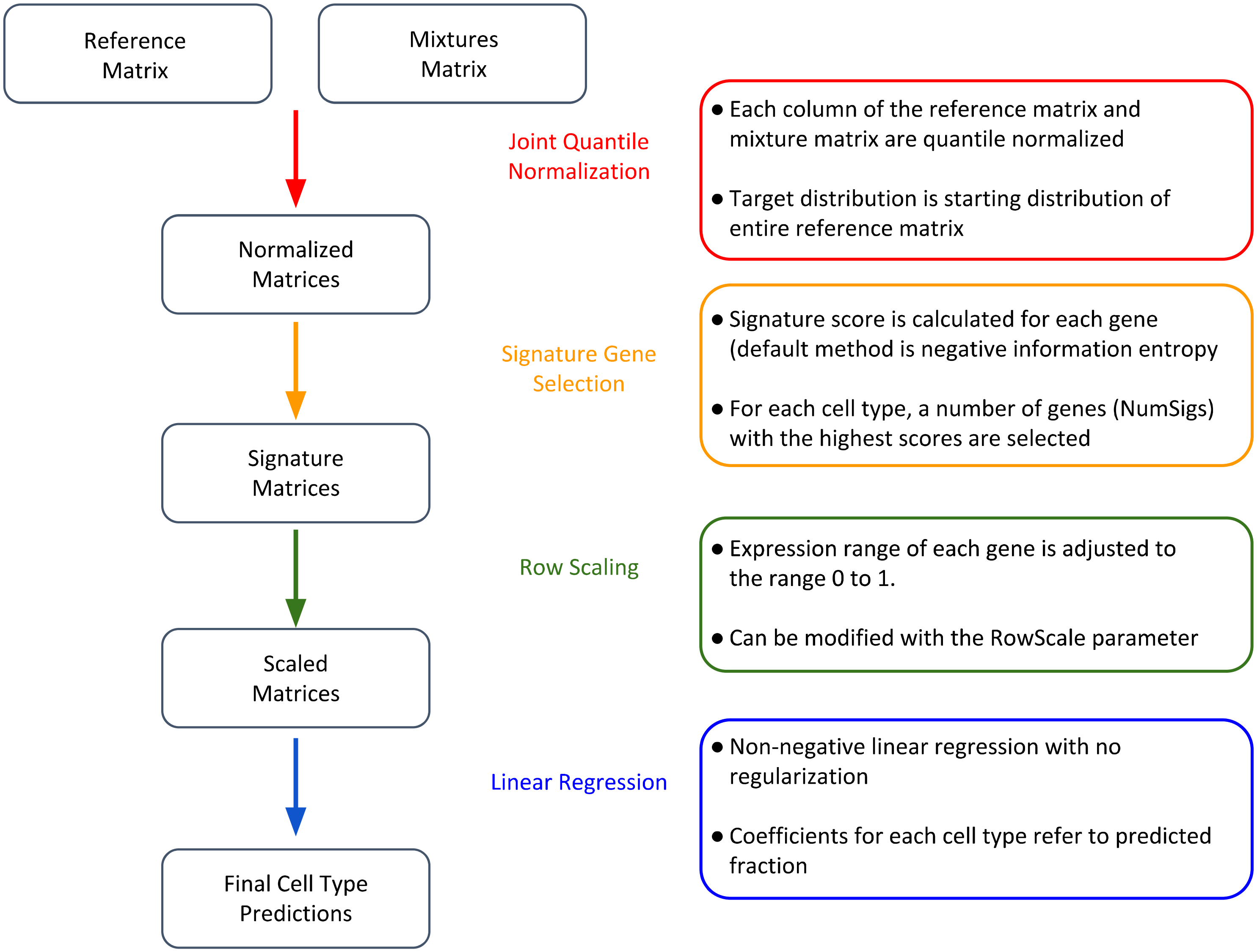
38. Carithers LJ, Ardlie K, Barcus M, Branton PA, Britton A, Buia SA, et al. A novel approach to high-quality postmortem tissue procurement: the GTEx project. Biopreserv Biobank. 2015;13: 311–319.

39. Inkeles MS, Scumpia PO, Swindell WR, Lopez D, Teles RMB, Graeber TG, et al. Comparison of molecular signatures from multiple skin diseases identifies mechanisms of immunopathogenesis. J Invest Dermatol. 2015;135: 151–159.

40. Nadel B. GEDIT. 2020. GitHub Repository: https://github.com/BNadel/GEDIT

41. Nadel BB; Lopez D; Montoya DJ; Ma F; Waddel H; Khan MM; Mangul S; Pellegrini M. Supporting data for "The Gene Expression Deconvolution Interactive Tool (GEDIT): Accurate Cell Type Quantification from Gene Expression Data" *GigaScience* Database 2020. http://dx.doi.org/10.5524/100846

Figure 1



- Each column of the reference matrix and mixture matrix are quantile normalized
- Target distribution is starting distribution of entire reference matrix

- Signature score is calculated for each gene (default method is negative information entropy
- For each cell type, a number of genes (NumSigs) with the highest scores are selected

- Expression range of each gene is adjusted to the range 0 to 1.
- Can be modified with the RowScale parameter

- Non-negative linear regression with no regularization
- Coefficients for each cell type refer to predicted fraction

Figure 2

Click here to access/download;Figure;Figure 2.pdf

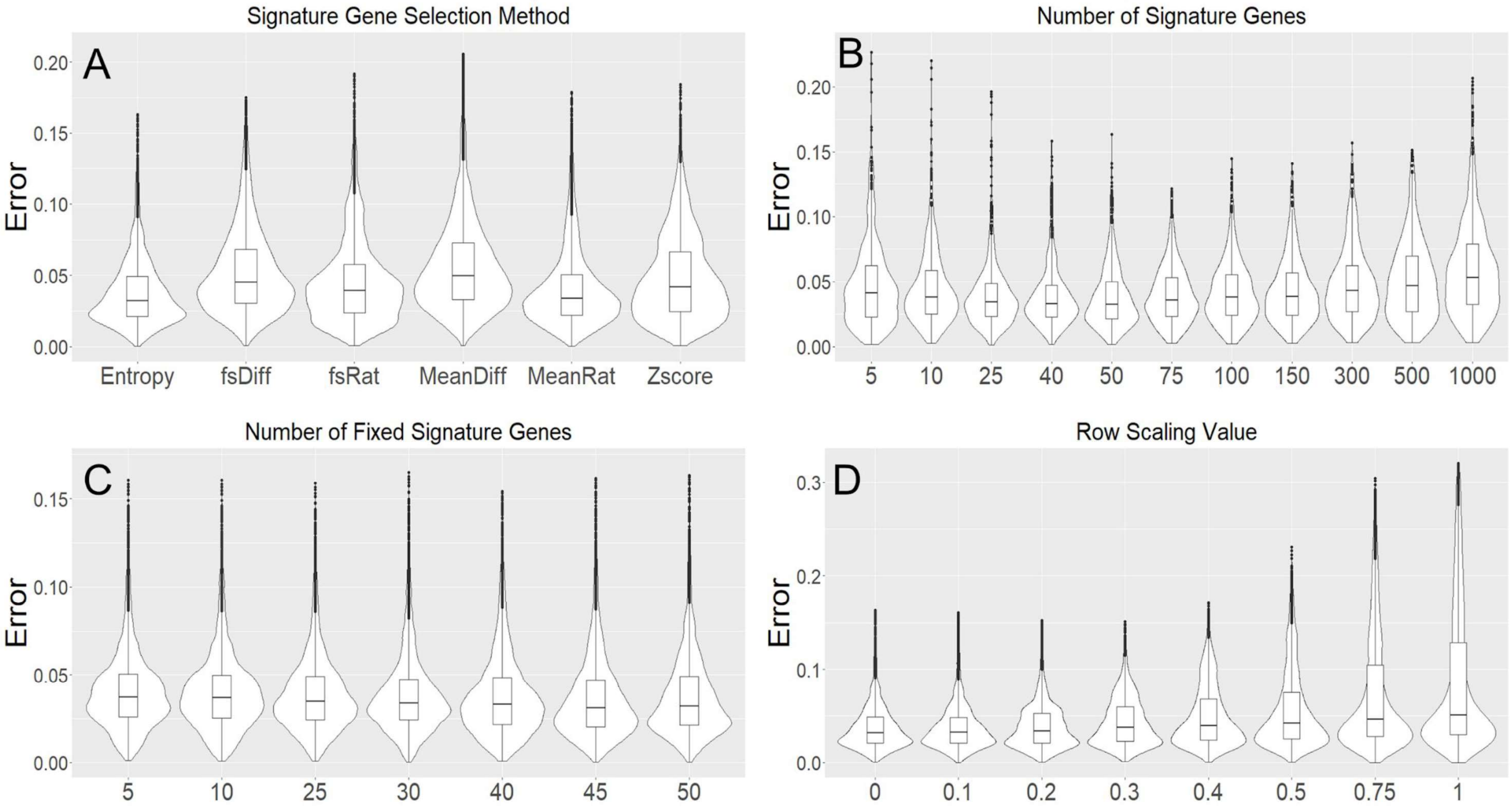


Figure 3

Cell Type	Monocytes	Neutrophils	B Cells	NK Cells	CD4+ T cells	Macrophages
CD14	338.4	163.9	18.9	16.9	19.2	105.9
THEMIS	9.7	11.6	8.4	13.2	52.0	8.7

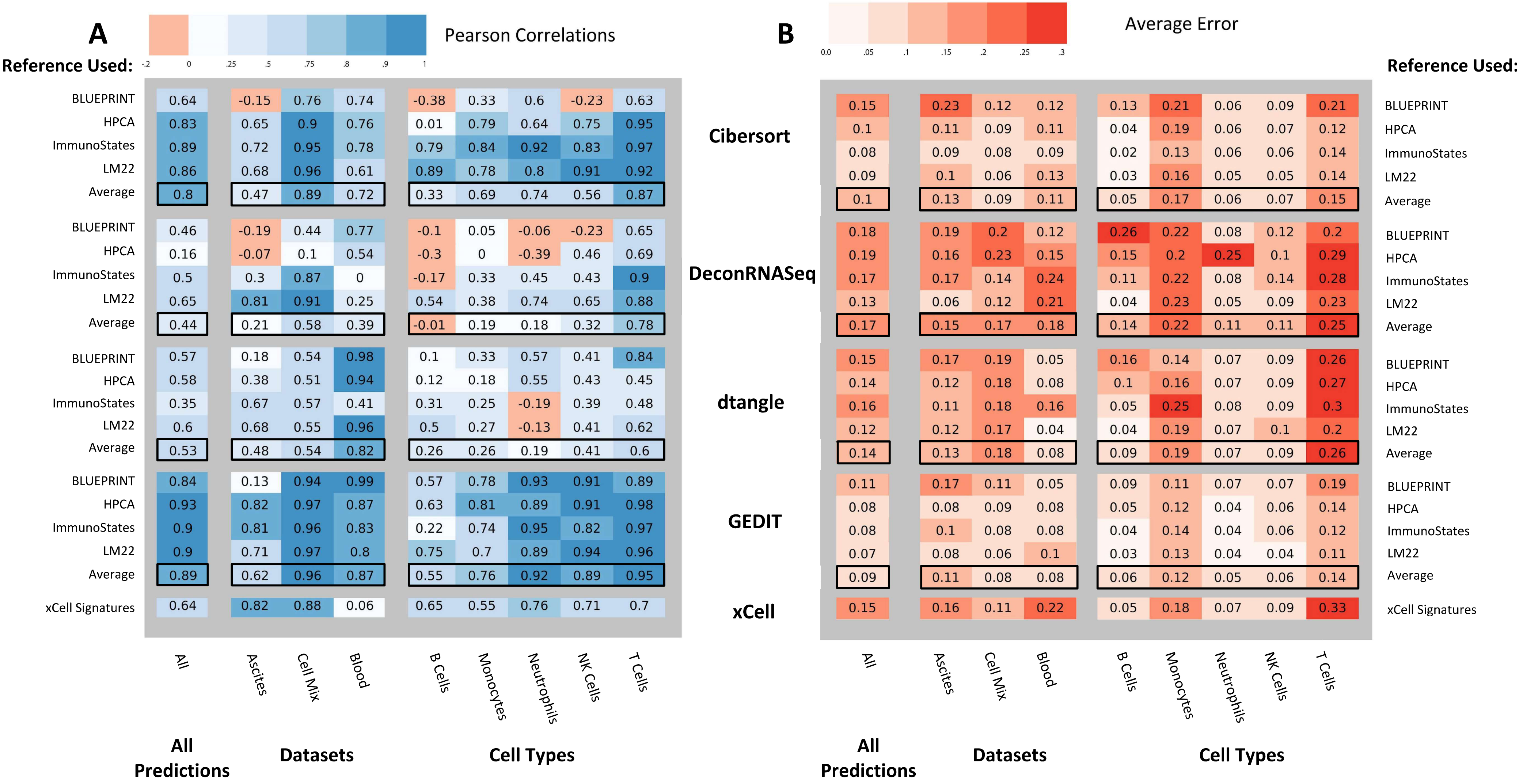
Mixture 1	Mixture 2
22.3	95.0
50.3	20.3

Row Scaling Transformation

Cell Type	Monocytes	Neutrophils	B Cells	NK Cells	CD4+ T cells	Macrophages
CD14	1.0	.46	.01	0.0	.01	.28
THEMIS	.03	.07	0.0	.11	1.0	.01

Mixture 1	Mixture 2
.02	.24
.96	.27

Figure 4



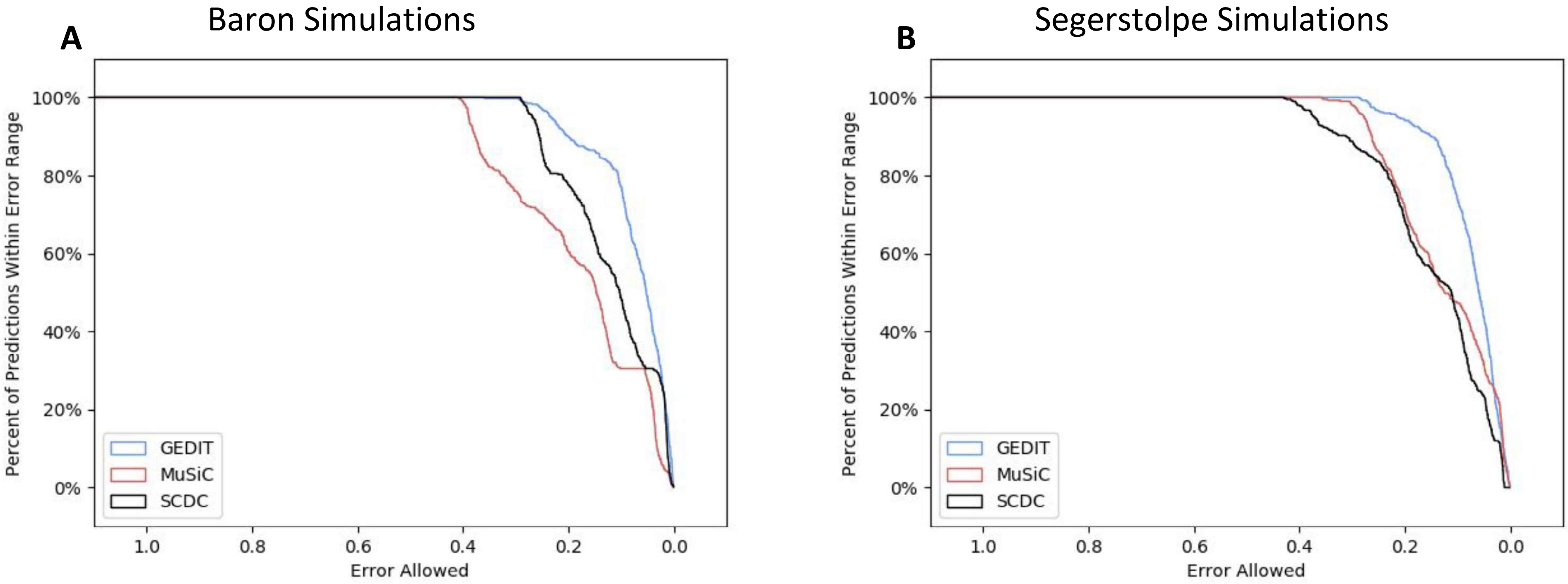
A

Method	Reference Used	All	Ascites	Cell Mix	Blood	B Cells	Monocytes	Neutrophils	NK Cells	T Cells
Cibersort	BLUEPRINT	0.64	-0.15	0.76	0.74	-0.38	0.33	0.6	-0.23	0.63
	HPCA	0.83	0.65	0.9	0.76	0.01	0.79	0.64	0.75	0.95
	ImmunoStates	0.89	0.72	0.95	0.78	0.79	0.84	0.92	0.83	0.97
	LM22	0.86	0.68	0.96	0.61	0.89	0.78	0.8	0.91	0.92
	Average	0.8	0.47	0.89	0.72	0.33	0.69	0.74	0.56	0.87
DeconRNASeq	BLUEPRINT	0.46	-0.19	0.44	0.77	-0.1	0.05	-0.06	-0.23	0.65
	HPCA	0.16	-0.07	0.1	0.54	-0.3	0	-0.39	0.46	0.69
	ImmunoStates	0.5	0.3	0.87	0	-0.17	0.33	0.45	0.43	0.9
	LM22	0.65	0.81	0.91	0.25	0.54	0.38	0.74	0.65	0.88
	Average	0.44	0.21	0.58	0.39	-0.01	0.19	0.18	0.32	0.78
dtangle	BLUEPRINT	0.57	0.18	0.54	0.98	0.1	0.33	0.57	0.41	0.84
	HPCA	0.58	0.38	0.51	0.94	0.12	0.18	0.55	0.43	0.45
	ImmunoStates	0.35	0.67	0.57	0.41	0.31	0.25	-0.19	0.39	0.48
	LM22	0.6	0.68	0.55	0.96	0.5	0.27	-0.13	0.41	0.62
	Average	0.53	0.48	0.54	0.82	0.26	0.26	0.19	0.41	0.6
GEDIT	BLUEPRINT	0.84	0.13	0.94	0.99	0.57	0.78	0.93	0.91	0.89
	HPCA	0.93	0.82	0.97	0.87	0.63	0.81	0.89	0.91	0.98
	ImmunoStates	0.9	0.81	0.96	0.83	0.22	0.74	0.95	0.82	0.97
	LM22	0.9	0.71	0.97	0.8	0.75	0.7	0.89	0.94	0.96
	Average	0.89	0.62	0.96	0.87	0.55	0.76	0.92	0.89	0.95
xCell	xCell Signatures	0.64	0.82	0.88	0.06	0.65	0.55	0.76	0.71	0.7

B

Method	Reference Used	All	Ascites	Cell Mix	Blood	B Cells	Monocytes	Neutrophils	NK Cells	T Cells
Cibersort	BLUEPRINT	0.15	0.23	0.12	0.12	0.13	0.21	0.06	0.09	0.21
	HPCA	0.1	0.11	0.09	0.11	0.04	0.19	0.06	0.07	0.12
	ImmunoStates	0.08	0.09	0.08	0.09	0.02	0.13	0.06	0.06	0.14
	LM22	0.09	0.1	0.06	0.13	0.03	0.16	0.05	0.05	0.14
	Average	0.1	0.13	0.09	0.11	0.05	0.17	0.06	0.07	0.15
DeconRNASeq	BLUEPRINT	0.18	0.19	0.2	0.12	0.26	0.22	0.08	0.12	0.2
	HPCA	0.19	0.16	0.23	0.15	0.15	0.2	0.25	0.1	0.29
	ImmunoStates	0.17	0.17	0.14	0.24	0.11	0.22	0.08	0.14	0.28
	LM22	0.13	0.06	0.12	0.21	0.04	0.23	0.05	0.09	0.23
	Average	0.17	0.15	0.17	0.18	0.14	0.22	0.11	0.11	0.25
dtangle	BLUEPRINT	0.15	0.17	0.19	0.05	0.16	0.14	0.07	0.09	0.26
	HPCA	0.14	0.12	0.18	0.08	0.1	0.16	0.07	0.09	0.27
	ImmunoStates	0.16	0.11	0.18	0.16	0.05	0.25	0.08	0.09	0.3
	LM22	0.12	0.12	0.17	0.04	0.04	0.19	0.07	0.1	0.2
	Average	0.14	0.13	0.18	0.08	0.09	0.19	0.07	0.09	0.26
GEDIT	BLUEPRINT	0.11	0.17	0.11	0.05	0.09	0.11	0.07	0.07	0.19
	HPCA	0.08	0.08	0.09	0.08	0.05	0.12	0.04	0.06	0.14
	ImmunoStates	0.08	0.1	0.08	0.08	0.04	0.14	0.04	0.06	0.12
	LM22	0.07	0.08	0.06	0.1	0.03	0.13	0.04	0.04	0.11
	Average	0.09	0.11	0.08	0.08	0.06	0.12	0.05	0.06	0.14
xCell	xCell Signatures	0.15	0.16	0.11	0.22	0.05	0.18	0.07	0.09	0.33

Figure 5



C

	Baron Simulations		Segerstolpe Simulations	
Tool	**Correlation**	**Average Error**	**Correlation**	**Average Error**
GEDIT	.870	.073	.912	.074
MuSiC	.494	.177	.92	.137
SCDC	.889	.118	.94	.149

Figure 6

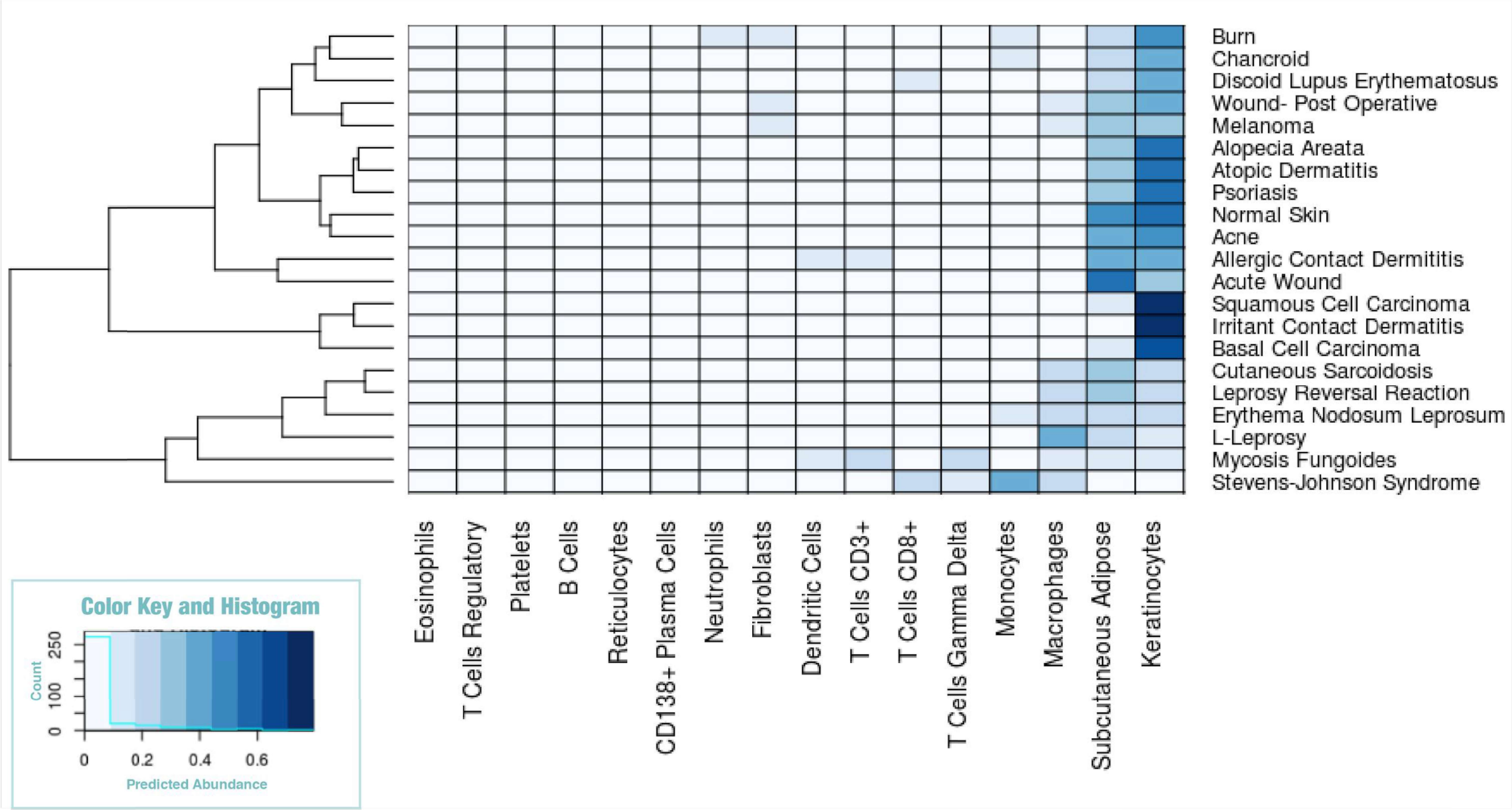
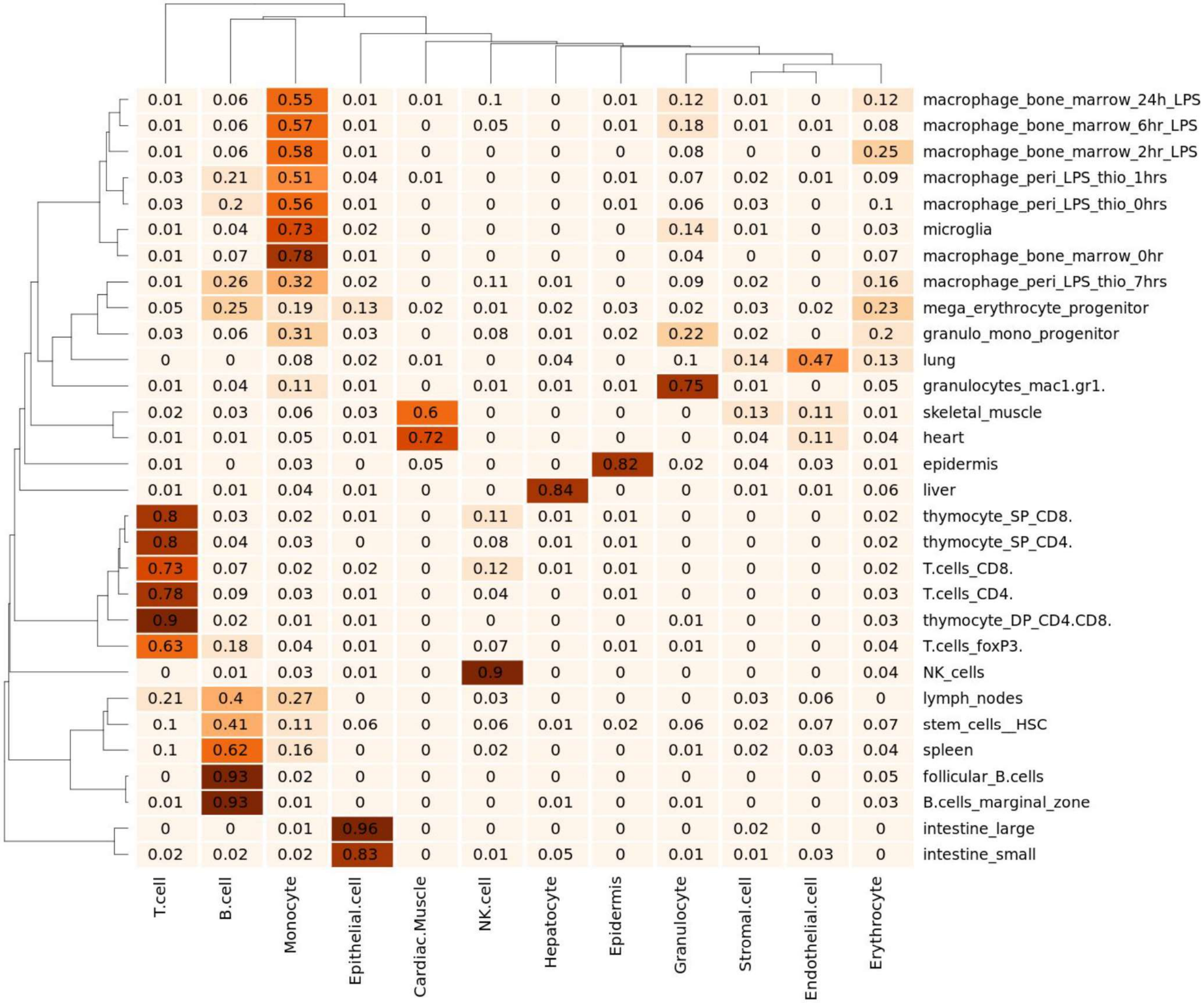
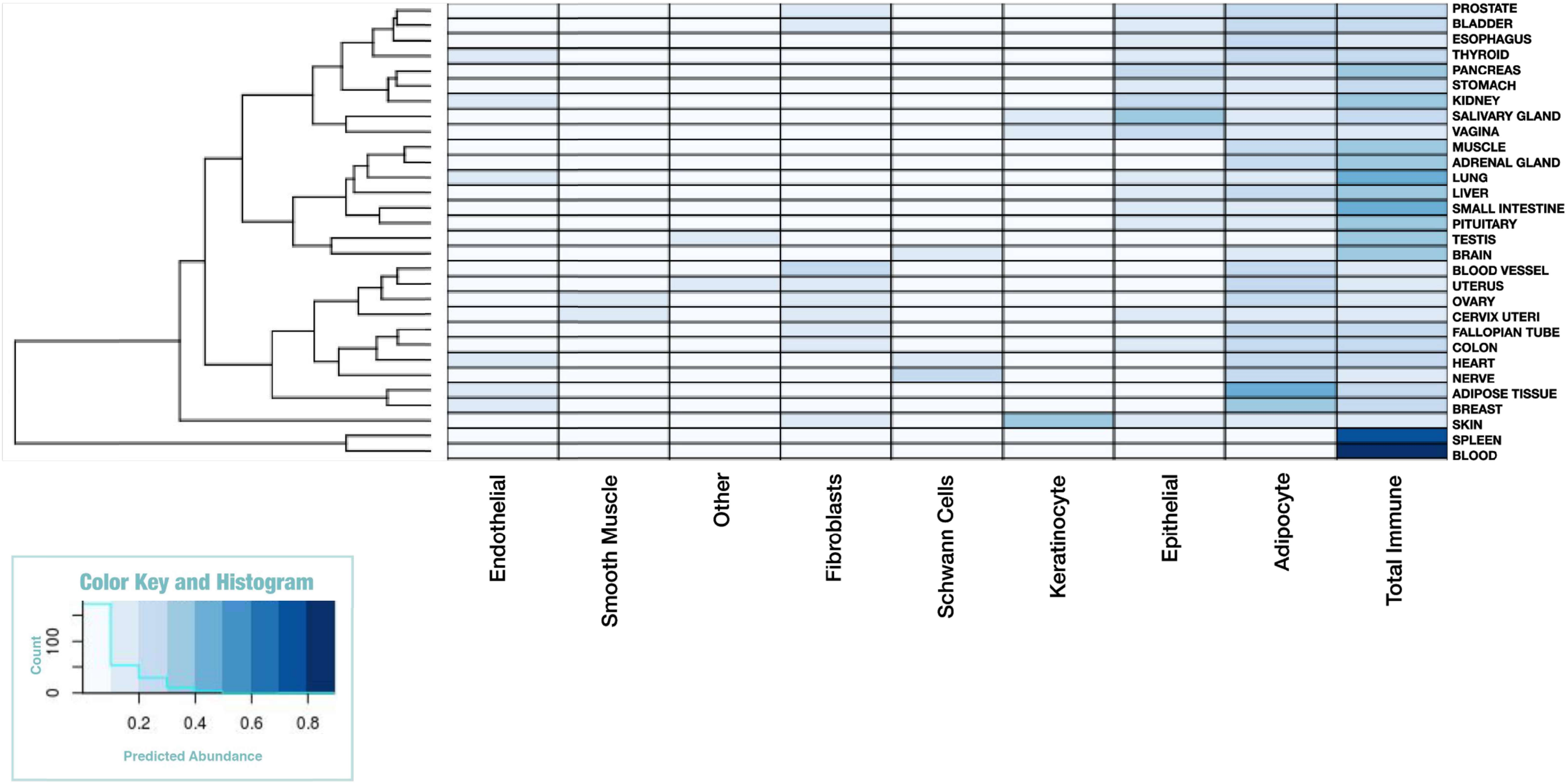

Burn
Chancroid
Discoid Lupus Erythematosus
Wound- Post Operative
Melanoma
Alopecia Areata
Atopic Dermatitis
Psoriasis
Normal Skin
Acne
Allergic Contact Dermitit is
Acute Wound
Squamous Cell Carcinoma
Irritant Contact Dermatitis
Basal Cell Carcinoma
Cutaneous Sarcoidosis
Leprosy Reversal Reaction
Erythema Nodosum Leprosum
L-Leprosy
Mycosis Fungoides
Stevens-Johnson Syndrome
Eosinophils
T Cells Regulatory
Platelets
B Cells
Reticulocytes
CD138+ Plasma Cells
Neutrophils
Fibroblasts
Dendritic Cells
T Cells CD3+
T Cells CD8+
T Cells Gamma Delta
Monocytes
Macrophages
Subcutaneous Adipose
Keratinocytes
Color Key and Histogram
Count
0
100
250
0
0.2
0.4
0.6
Predicted Abundance

Figure 7



	T.cell	B.cell	Monocyte	Epithelial.cell	Cardiac.Muscle	NK.cell	Hepatocyte	Epidermis	Granulocyte	Stromal.cell	Endothelial.cell	Erythrocyte
macrophage_bone_marrow_24h_LPS	0.01	0.06	0.55	0.01	0.01	0.1	0	0.01	0.12	0.01	0	0.12
macrophage_bone_marrow_6hr_LPS	0.01	0.06	0.57	0.01	0	0.05	0	0.01	0.18	0.01	0.01	0.08
macrophage_bone_marrow_2hr_LPS	0.01	0.06	0.58	0.01	0	0	0	0	0.08	0	0	0.25
macrophage_peri_LPS_thio_1hrs	0.03	0.21	0.51	0.04	0.01	0	0	0.01	0.07	0.02	0.01	0.09
macrophage_peri_LPS_thio_0hrs	0.03	0.2	0.56	0.01	0	0	0	0.01	0.06	0.03	0	0.1
microglia	0.01	0.04	0.73	0.02	0	0	0	0	0.14	0.01	0	0.03
macrophage_bone_marrow_0hr	0.01	0.07	0.78	0.01	0	0	0	0	0.04	0	0	0.07
macrophage_peri_LPS_thio_7hrs	0.01	0.26	0.32	0.02	0	0.11	0.01	0	0.09	0.02	0	0.16
mega_erythrocyte_progenitor	0.05	0.25	0.19	0.13	0.02	0.01	0.02	0.03	0.02	0.03	0.02	0.23
granulo_mono_progenitor	0.03	0.06	0.31	0.03	0	0.08	0.01	0.02	0.22	0.02	0	0.2
lung	0	0	0.08	0.02	0.01	0	0.04	0	0.1	0.14	0.47	0.13
granulocytes_mac1.gr1.	0.01	0.04	0.11	0.01	0	0.01	0.01	0.01	0.75	0.01	0	0.05
skeletal_muscle	0.02	0.03	0.06	0.03	0.6	0	0	0	0	0.13	0.11	0.01
heart	0.01	0.01	0.05	0.01	0.72	0	0	0	0	0.04	0.11	0.04
epidermis	0.01	0	0.03	0	0.05	0	0	0.82	0.02	0.04	0.03	0.01
liver	0.01	0.01	0.04	0.01	0	0	0.84	0	0	0.01	0.01	0.06
thymocyte_SP_CD8.	0.8	0.03	0.02	0.01	0	0.11	0.01	0.01	0	0	0	0.02
thymocyte_SP_CD4.	0.8	0.04	0.03	0	0	0.08	0.01	0.01	0	0	0	0.02
T.cells_CD8.	0.73	0.07	0.02	0.02	0	0.12	0.01	0.01	0	0	0	0.02
T.cells_CD4.	0.78	0.09	0.03	0.01	0	0.04	0	0.01	0	0	0	0.03
thymocyte_DP_CD4.CD8.	0.9	0.02	0.01	0.01	0	0	0	0	0.01	0	0	0.03
T.cells_foxP3.	0.63	0.18	0.04	0.01	0	0.07	0	0.01	0.01	0	0	0.04
NK_cells	0	0.01	0.03	0.01	0	0.9	0	0	0	0	0	0.04
lymph_nodes	0.21	0.4	0.27	0	0	0.03	0	0	0	0.03	0.06	0
stem_cells__HSC	0.1	0.41	0.11	0.06	0	0.06	0.01	0.02	0.06	0.02	0.07	0.07
spleen	0.1	0.62	0.16	0	0	0.02	0	0	0.01	0.02	0.03	0.04
follicular_B.cells	0	0.93	0.02	0	0	0	0	0	0	0	0	0.05
B.cells_marginal_zone	0.01	0.93	0.01	0	0	0	0.01	0	0.01	0	0	0.03
intestine_large	0	0	0.01	0.96	0	0	0	0	0	0.02	0	0
intestine_small	0.02	0.02	0.02	0.83	0	0.01	0.05	0	0.01	0.01	0.03	0

Figure 8


PROSTATE
BLADDER
ESOPHAGUS
THYROID
PANCREAS
STOMACH
KIDNEY
SALIVARY GLAND
VAGINA
MUSCLE
ADRENAL GLAND
LUNG
LIVER
SMALL INTESTINE
PITUITARY
TESTIS
BRAIN
BLOOD VESSEL
UTERUS
OVARY
CERVIX UTERI
FALLOPIAN TUBE
COLON
HEART
NERVE
ADIPOSE TISSUE
BREAST
SKIN
SPLEEN
BLOOD
Endothelial
Smooth Muscle
Other
Fibroblasts
Schwann Cells
Keratinocyte
Epithelial
Adipocyte
Total Immune
Color Key and Histogram
Count
0
100
0.2
0.4
0.6
0.8
Predicted Abundance

Dear Giga Science Reviewer

Reviewer reports:

Reviewer #1: The authors present an algorithm called GEDIT using information from a reference dataset to estimate cell type abundances in a target dataset. This manuscript is not qualified to be published in the Giga Science because of the following reasons:
1. GEDIT does not show enough novelty.

As the authors stated in the response document, GEDIT has two key innovations: signature gene selection by information entropy and the row scaling step. As shown in Figure 2a, the signature gene selection only has a limited improvement than the others in terms of error. The authors also did not have enough evidence to support how and why the row scaling step is helpful. On top of these 2 data preprocessing steps, I did not find any innovations on the model of a non-negative linear regression.

We refer the reviewer to figure 4D, in which we demonstrate that row scaling dramatically improves accuracy of predictions. Here, we vary the row scaling parameter in order to evaluate the effect of row scaling on accuracy. When this parameter is set to 0 (the default value for GEDIT) both average and maximum error are greatly reduced compared to a setting of 1.0 (i.e. row scaling is disabled).

2. GEDIT does not add much values in the field.
Although GEDIT is shown to have appealing results in comparison to existing methods on benchmarking experiments, it doesn't significantly outperform other methods in real data analysis in regards to Pearson correlation and average error. As demonstrated in Supplementary Figures 4 and 5, a main determinant of results quality for deconvolution is the reference used, not necessarily the algorithm.

We refer to Figure 4, which demonstrates that GEDIT does indeed outperform other tools in terms of Pearson correlation. When any of the four references are used, GEDIT produces higher correlations between predicted and actual fractions than any other tool (the leftmost “all” column). In addition, the highest observed correlations for each of the 3 mixtures sets (ascites, cellmix, blood) are achieved by GEDIT (i.e. when the Human Primary Cell Atlas, LM22, and BLUEPRINT are used as reference sources, respectively). Lastly, unlike the three other reference-based tools, GEDIT produces positive correlations for all cell types regardless of choice of reference. GEDIT similarly outperforms other tools when evaluated by error (Supplementary Figure 2).

Moreover, we include in the supplementary materials an additional comparison of bulk deconvolution tools (Supplementary Figure 6). Here, we compare the error of CIBERSORT, DeconRNASeq, dtangle, and GEDIT when applied to simulated pancreatic islet mixtures. Again, GEDIT outperforms the other deconvolution tools.

In addition, we are pursuing a separate project performing extensive benchmarking of the current field of deconvolution tools. We believe more comprehensive comparisons of these tools’ performances is appropriate, but that such an undertaking represents a separate project that should not be bundled with the publication of a new tool. For the interest of the editors and reviewers, we attach an early version of this benchmarking manuscript.

3. I still think comparing GEDIT to other methods using single-cell RNA-seq as a reference is necessary.
Utilizing single-cell RNA-seq for deconvolution becomes cutting-edge research and many tools have been designed for this purpose such as MuSic. These methods are proved to have superior performance using microarray as references and are commonly used. It is critical to demonstrate whether GEDIT has better performance than these methods.

At the reviewer’s request, we have performed a comparison between GEDIT and two well known single cell deconvolution tools (SCDC and MuSiC). Specifically, we utilize the testing framework developed by the SCDC authors to prepare synthetic mixtures using single cell data from pancreatic islets. We include an additional section describing these results in the main manuscript (text below), and also refer the reviewer to Figure 5 and Supplementary Figure 5.

Comparison to Single Cell Methods

We also compare GEDIT to two contemporary deconvolution tools that utilize single cell data as their reference, namely SCDC and MuSiC [10,11]. We reproduce the steps provided by the SCDC authors to generate two sets of 100 simulated pancreatic mixtures. These data are created *in silico* using single cell data from two recent studies, and contain randomized mixtures of alpha, beta, delta, and gamma cells from pancreatic islets [33,34]. Data from a third study was used as a reference for all 3 tools, and similarly contains alpha, beta, gamma, and delta cells [35]. In the case of SCDC and MuSiC, these data are used in their original single cell form. For GEDIT, pseudo-bulk expression profiles for each of the four cell types were created by averaging the expression values of each member cell (e.g. expression of all alpha cells were averaged to create an alpha cell reference profile).

The results of GEDIT compare favorably to the two single cell tools (Figure 5). GEDIT produces the lowest error on the two sets of simulated mixtures by a significant margin. Based on the metric of correlation between predicted and actual fractions, GEDIT produces results comparable to SCDC, and either comparable or superior to MuSiC, depending on the set of mixtures (Table 4, Supplementary Figure 5). Thus, by using the methodology of averaging cell clusters in the reference dataset, GEDIT can be applied to datasets suitable for SCDC or MuSiC. We also apply three other bulk deconvolution tools to this same dataset, and show that GEDIT provides the best performance out of the four (Supplementary Figure 6).

Reviewer #2: For the most part the responses are sufficient and the authors have addressed the concerns, and the manuscript is improved. I especially appreciate that claims have been toned down and better contextualized.

A small issue remains about minor comment 6. My point was that the readers should be made clearly aware that doing three deconvolutions is not ideal, and strictly speaking invalid (e.g. cell contents totalling over 100%). I suggested two hypothetical ways to avoid this and am not at all surprised it's not easy to fix with data on hand. In the context of the demonstration in this particular study, what the authors did originally is acceptable. The problem I raised is if people start copying that practice in their own studies, and in the authors' own interest they presumably wouldn't want to be seen as endorsing it. The statement in the supplement that "creating a comprehensive reference from single cell data will likely produce superior results" should be more prominent, and it's not just about superior results, it's also about validity of having a single reference vs. multiple independent deconvolutions.

To be concrete, I'd suggest that in the main body a parenthetical could be added to the effect that "it would be more appropriate to have a single reference containing all cell types and performing a single deconvolution; see supplement for discussion". As it stands I don't think the addition to the supplement is referenced in the main paper.

We appreciate this feedback. We have adjusted the language in this section as follows:

> To assess the use of GEDIT across very large datasets, we applied the tool to 17,382 GTEx RNA-seq samples collected from various tissues. **However, no single reference contained all cell types expected to be present and combining references from separate experiments and platforms is problematic (Supplementary Figures 9-11).** Therefore, we took an alternate approach by performing deconvolution three times using three separate references (BlueCode, Human Primary Cell Atlas, Skin Signatures). We then combine these outputs by taking their median value; after normalization, we treat this median value as a final cell type estimate (see Supplementary Materials for more details). While this approach did enable predictions spanning a larger number of cell types than are present in any one reference matrix, **it must me noted that it is not a proper substitute for a single unified reference** (Figure 8).